

Study the combination of brain MRI imaging and other datatypes to improve Alzheimer's disease diagnosis

Jonathan Stubblefield^{1,2,3*}, Alan Kronberger^{1,4*}, Jason Causey^{1,3}, Jake Qualls^{1,3}, Jennifer Fowler^{1,4}, Kaiman Zeng⁵, Karl Walker⁶, Xiuzhen Huang^{1,7§} and for the Alzheimer's Disease Neuroimaging Initiative[†]

¹Center for No-Boundary Thinking (CNBT), Arkansas State University, Jonesboro, AR

²Arkansas Biosciences Institute, Arkansas State University, Jonesboro, AR

³Department of Computer Science, Arkansas State University, Jonesboro, AR

⁴Molecular Biosciences Graduate Program, Arkansas State University, Jonesboro, AR

⁵Electrical Engineering, Arkansas Tech University, Russellville, AR

⁶Department of Computer Science and Mathematics, University of Arkansas at Pine Bluff, AR

⁷Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA

*The authors JS, AK are joint first authors. §Correspondence to: Xiuzhen.Huang@cshs.org; All authors are with the Arkansas AI-Campus, and the Center for No-Boundary Thinking (CNBT).

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract.

Alzheimer's Disease (AD) is a degenerative brain disease and is the most common cause of dementia. Despite being a common disease, AD is poorly understood. Current medical treatments for AD are aimed at slowing the progression of the disease. So early detection of AD is important to intervene at an early stage of the disease. In recent years, by using machine learning predictive algorithms, assisted clinic diagnosis has received great attention due to its success of machine learning advances in the domains of computer vision. In this study, we have combined brain MRI imaging features and the features of other datatypes, and adopted various models, including XGBoost, logistic regression, and k-Nearest Neighbors, to improve AD diagnosis. We evaluated the models on the benchmark dataset of Alzheimer's Disease Neuroimaging Initiative. Experiment results show that the logistic regression model is the top performer in terms of evaluation metrics of precision, recall, and F1-score. The prediction of the models could provide valuable information for diagnosis and prognosis of patients with suspected Alzheimer's disease. The XGBoost model achieves a comparable performance and has the potential to serve as a valuable diagnostic tool for patients with suspected AD with its self-validation by re-discovering previously known associations with AD.

1. Background

Alzheimer's disease (AD) is a degenerative brain disease and is the number one cause of dementia [1]. It is estimated that 60% - 80% of dementia cases are due to AD or have shown Alzheimer's disease brain changes with one or more other causes of dementia, such as Cerebrovascular disease or Lewy body disease [2]. AD typically manifests with impairment in memory, judgement, executive function, cognitive decline, and behavioral and psychological changes [1]. Despite being a common disease, AD is poorly understood [1]. The cause and pathogenesis of AD remain open areas of research [1]. As of yet, there is no cure for AD [1]. Current medical treatments for AD are aimed at slowing the progression of the disease [1].

1.1 Current Diagnostics

The only definitive diagnosis for AD is a brain biopsy with a histopathological examination, which is usually only done during an autopsy [1]. This makes the diagnosis of AD very challenging. In practice, AD is usually diagnosed by a physician's clinical assessment of the patient [1]. Neuroimaging, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), can be used to search for alternative diagnoses or document the progression of the disease [1]. Biomarkers for AD have been studied and can support a diagnosis of AD but are not currently recommended as diagnostic tests [1].

In recent years, data-driven approaches have received great attention due to the advancement of big data, computational power, and machine learning algorithms. Because the AD related brain changes is a slowly developing process and could begin years before symptoms emerge [2], early detection of AD has a great potential to provide early intervention, slow disease progression, and reduce the care cost.

1.2 Related Work

Machine learning algorithms have been applied in the medical field to assist diagnosis for a lot of diseases, such as heart disease, Parkinson's disease, and Coronavirus disease. A great amount of research has been performed on using machine learning algorithms to assist AD diagnosis, which trains a classification model to predict if a patient will have AD, develop mild cognitive impairment (MCI), or remain normal (represented as normal control, NC). The studied machine learning algorithms can be categorized as classical machine learning models and deep learning models. The classical machine learning models mainly include logistic regression, support vector machine (SVM), and tree-based models. The deep learning models are different architectures of neural networks.

Antor et. al. performed a comparative analysis of logistic regression, decision tree, random forest, and SVM models to predict AD. They also evaluated the models with fine-tuning to those without fine-tune. Their experiment results showed that the SVM best performed in detecting dementia in terms of accuracy, recall, precision, AUC, and F1 score. They claimed a more complex model such as random forest classifier suffered from overfitting.

Grueso and Viejo-Sobera [3] conducted a comprehensive review and qualitative analysis over 116 studies selected from 452 publications that used machine learning to predict progression of AD. It was found that the SVM was the most commonly used a classifier with a mean accuracy of 75.4%. The SVM algorithm uses a kernel function to map the feature space to a higher

dimension and usually a polynomial or Gaussian function was used [Automated]. Ensembling models of SVM and other methods have also been proposed to optimize the feature selection and to improve the generalization of models [Thung].

Deep learning techniques have also gained great attentions in recent years due to its success in the domains of computer vision and natural language processing. Saleem [DeepLearning] et. al. have reviewed a number of studies for AD diagnosis based on the major deep learning models, including feed-forward deep neural networks (DNN), convolutional neural networks (CNN), auto-encoder (AE), recurrent neural networks (RNN), deep belief network (DBN), generative adversarial networks (RNN), and hybrid models. The authors observed deep learning techniques in general outperform classical machine learning techniques, with accuracies ranging from 53.17% to 99.95% for classifications of NC/AD, NC/AD, MCI/AD or NC/MCI/AD. However, most deep learning models lack explainability and as a result lead to criticisms for being adopted into clinical practice. Paper [Explainability] discussed the role of explainable AI from the perspectives of technology, medicine, legislation, ethics, and societies. They argued that emitting explainability in clinical decision system would jeopardize the core ethic value of medicine and public health.

1.3 Alzheimer's Disease Neuroimaging Initiative (ADNI)

To combat this current lack of usable diagnostic studies the ADNI study was started in 2004 [about_adni]. This large, longitudinal, multi-center study has accumulated data on a large number of patients with data clinical, MRI imaging, and proteomic data [about_adni].

2. Methods

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

In this paper, we present a set of models developed with data from the ADNI study for the diagnosis of AD using proteomic and MRI imaging data.

2.1 ADNI Patients

We used the ADNI dataset for this study. We sampled 167 individuals across the ADNI 1 and ADNI 2 datasets that had both proteomic and MRI imaging data. These individuals included a mix of patients with AD, mild-cognitive impairment, and healthy controls.

2.2 ADNI Imaging Features

We used the ADNI imaging features that had already been processed by the study using FreeSurfer. FreeSurfer is an open-source software package used to visualize and analyze MRI images. The MRI data of the ADNI 1 dataset was obtained using FS V4.4. FS V5.1 was used to process the ADNI 2 dataset

[@adni_methods]. The results contained 70 cortical thickness measurements and 42 cortical volume measurements [@adni_methods].

2.3 ADNI Proteomic Features

The ADNI proteomic features consist of protein levels from samples of the patients' cerebrospinal fluid. These samples were processed by Rules-Based Medicine using the Luminex xMAP [@biomarkers]. The data contained a total of 229 proteomic measurements.

2.4 Feature Normalization

All features were scaled element-wise using Scikit-Learn version 0.24.1's MinMaxScaler [@Pedregosa].

2.5 Different methods

XGBoost Method. We used XGBoost as our classification model during this project [@Chen]. We recorded results for each pairwise binary classification problem for the three classes, normal control (NC), mild-cognitive impairment (MCI), and Alzheimer's Disease (AD). We also recorded results for XGBoost classifying all three classes simultaneously. We ran the model in Python using Python 3.8.8 and the 1.3.3 version of the XGBoost package. The experiment was performed using 5-fold cross-validation with Scikit-Learn's stratified k-fold [@Pedregosa]. We used the 0.24.1 version of Scikit-Learn [@Pedregosa]. XGBoost's performance was evaluated on each datatype individually and both data types combined. The two datatypes were combined using simple concatenation of the feature arrays.

Logistic Regression Method. Logistic regression was chosen as a comparison model to XGBoost. Functioning as linear model predicting class probabilities, we expected it to approach the problem in very different than XGBoost's gradient boosted trees [@Chen]. We used Scikit-Learn version 0.24.1's version of logistic regression with default parameters aside from a max iteration of 500 [@Pedregosa].

K-Nearest Neighbors Method. K-Nearest neighbors (KNN) was also chosen as a comparison model. KNN is a non-parametric model that can vary its flexibility by changing the value of K. We used Scikit-Learn version 0.24.1's version of KNN, evaluating it for K values 2, 5, and 15 [@Pedregosa].

Model Evaluation. Each method was evaluated using five-fold cross-validation on each binary classification problem consisting of two paired classes (control vs Alzheimer's, control vs mild cognitive impairment, and mild cognitive impairment vs Alzheimer's) as well as the 3-class classification problem. Recall, precision, and f1-score were calculated for each classification problem and each fold. The averages with respect to each class were reported.

SHAP Feature Importance. We used the SHAP TreeExplainer utility to evaluate and graph the relative importance of the imaging and proteomic features to the XGBoost model [@Lundberg_tree].

3. Results

Here we present the performance of the different methods with the MRI image features and proteomics data features respectively, and the combination of the two types of data features.

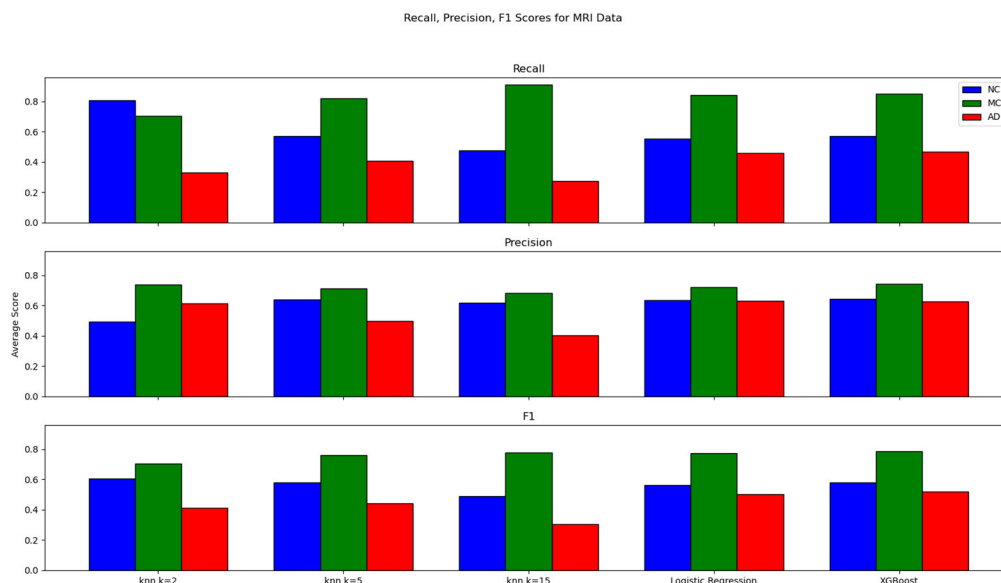


Figure 1: Recall, Precision, F1 Scores for MRI Data. These graphs show various classification metrics with respect to each class. From top to bottom, the graphs show recall, precision, and f1-score. Metrics for the NC class are displayed in blue, MCI in green, and AD in red. The metrics are the averages across all trials.

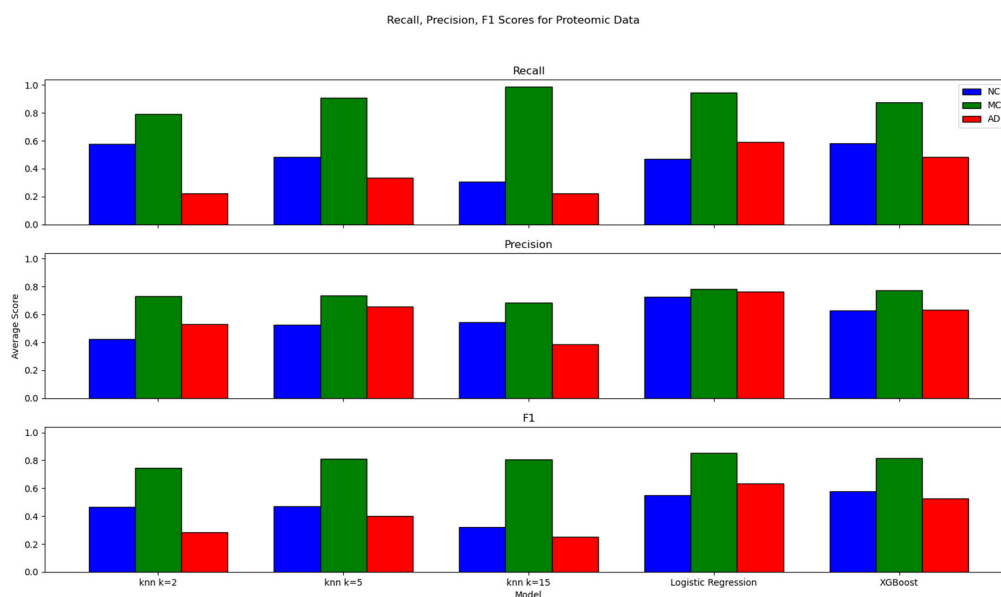


Figure 2: Recall, Precision, F1 Scores for Protein Data. These graphs show various classification metrics with respect to each class. From top to bottom, the graphs show recall, precision, and f1-score. Metrics for the NC class are displayed in blue, MCI in green, and AD in red. These metrics are the averages across all trials.

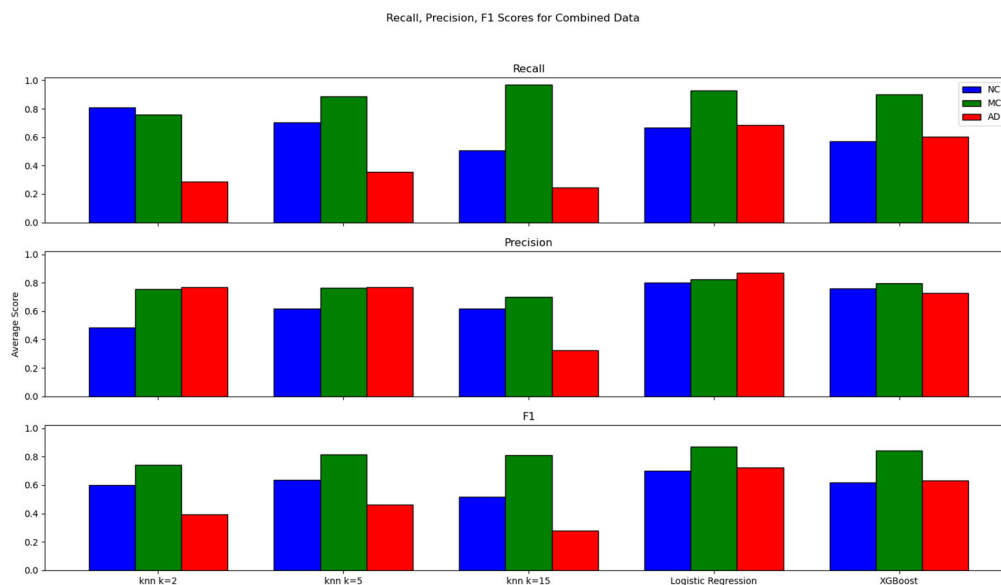


Figure 3: Recall, Precision, F1 Scores for Combined Data. These graphs show various classification metrics with respect to each class. From top to bottom, the graphs show recall, precision, and f1-score. Metrics for the NC class are displayed in blue, MCI in green, and AD in red. These metrics are the averages across all trials.

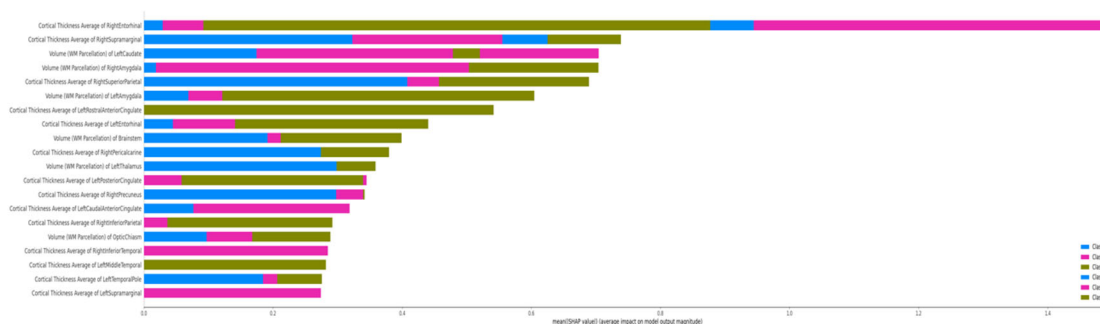


Figure 4: ADNI Summary Plot, Image Features. The overlapping bars show the average SHAP values of the top image features for each class.

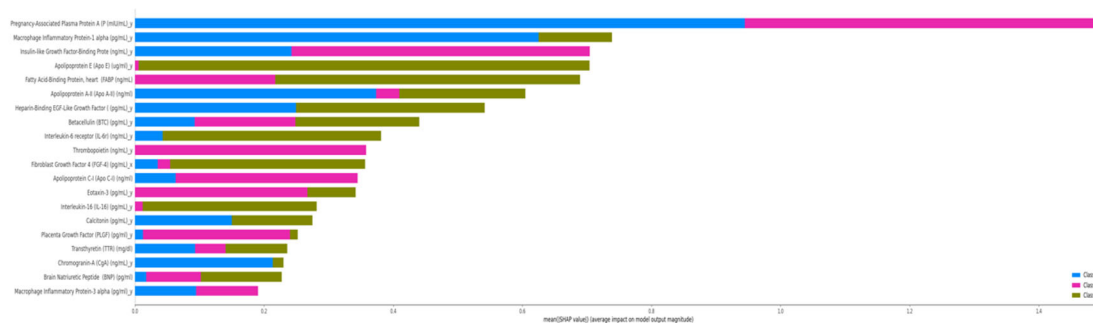


Figure 5: ADNI Summary Plot, Proteomic Features. The overlapping bars show the average SHAP values of the top proteomic features for each class.

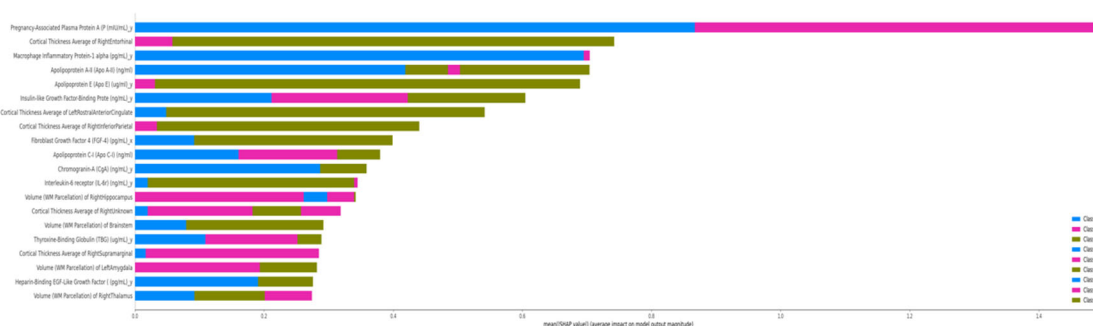


Figure 6: ADNI Summary Plot, All Features. The overlapping bars show the average SHAP values of the top imaging and proteomic features for each class.

Table 1: XGBoost performance on MRI Image Features Only. All scores are the average of 5-fold cross-validation. The recall, precision, and F1-scores for the trinary test (NCvMCIvAD) use a weighted average.

	Accuracy	Recall	Precision	F1
MCivAD	0.686894587	0.686894587	0.6580480014	0.6578067747
NCvMCI	0.788923077	0.7889230769	0.7550217483	0.7609002399
NCvAD	0.791428571	0.791428571	0.7142857143	0.7142857143
NCvMCIvAD	0.6237076649	0.6237076649	0.6237989168	0.604847908

Table 2: XGBoost Performance on Proteomic Features Only. All scores are the average of 5-fold cross-validation. The recall, precision, and F1-Scores for the trinary test (NCvMCIvAD) use a weighted average.

	Accuracy	Recall	Precision	F1
MCivAD	0.7612535613	0.7612535613	0.7602303998	0.7449376398
NCvMCI	0.8043076923	0.8043076923	0.8154784925	0.7897161275
NCvAD	0.7400000000	0.7400000000	0.7608008658	0.7339818077
NCvMCIvAD	0.6295900178	0.6295900178	0.5937355078	0.5894027955

Table 3: XGBoost Performance on MRI and Proteomic Features Combined. All scores are the average of 5-fold cross-validation. The recall, precision, and F1-Scores for the trinary test (NCvMCIvAD) use a weighted average.

	Accuracy	Recall	Precision	F1
MCivAD	0.8136752137	0.8136752137	0.8113215615	0.7942853606
NCvMCI	0.8043076923	0.8043076923	0.8154784925	0.7897161275
NCvAD	0.779047619	0.779047619	0.8011082251	0.7746031746
NCvMCIvAD	0.6896613191	0.6896613191	0.6984263775	0.6676843769

Table 4 shows the results improvement from combining across all experiments.

Table 4: Results Improvement. This table shows how well combining the two datatypes improves performance. The performances are in terms of F1-Score.

Problem	XGBoost Performance on Imaging Data	XGBoost Performance on Proteomic Data	XGBoost Performance on Combined Data	Performance Improvement
ADNI MCivAD	0.658	0.745	0.794	0.049
ADNI NCvMCI	0.761	0.790	0.790	0.000
ADNI NCvAD	0.714	0.734	0.775	0.041
ADNI Trinary	0.605	0.589	0.668	0.063

XGBoost displayed similar performance on both the imaging and proteomic datasets. For most of the experiments performed, combining the two kinds of data produced a synergistic increase in performance. The NCvAD experiment was an exception, showing no improvement when combining the two data sets. For the other three ADNI experiments, XGBoost produces an average improvement of 0.051 in terms of f1-score for the combined data.

The logistic regression model performed best out of all models tried for this experiment. Its performance is such that it may provide valuable information for diagnosis and prognosis of patients with suspected Alzheimer's disease.

With its reasonable performance and self-validation by re-discovering previously known associations with AD, the XGBoost model has the potential to serve as a valuable diagnostic tool for patients with suspected AD.

4. Discussions

4.1 Performance of the different Models

XGBoost: Overall, XGBoost produced good results on the ADNI dataset, producing a weighted f1-score of 0.76 on the combined data. XGBoost produced good synergy when combining the data, leading to an f1-score improvement of 0.05.

Logistic Regression: Logistic regression tended to outperform XGBoost in this analysis. Though they performed equally on MRI data alone, logistic regression outperformed XGBoost on the proteomic and combined data, yielding a final f1-score of 0.82 compared to XGBoost's 0.76. The logistic regression also showed more synergy between the datasets than XGBoost, improving by 0.07 when the datasets were combined.

K-Nearest Neighbors: Multiple values of K were tried for KNN. K=5 outperformed all other values of K on all datasets, although it still underperformed compared to XGBoost and logistic regression. The MRI dataset was less sensitive to changes in the value of K than the proteomic or combined datasets.

4.2 SHAP Feature Importance Analysis

The feature importance learned by XGBoost and reported using the SHAP utility matches up with expectations given current medical knowledge about AD. In MRI images, AD typically shows as non-specific generalized and focal atrophy of the brain [Wolk]. Hippocampal and medial temporal atrophy are the most characteristic findings of AD but remain insufficient for diagnosis by imaging alone [Wolk]. SHAP analysis of our XGBoost model showed that temporal lobe cortical thickness measurements were among the top features for the model using imaging features alone. Additionally, at least one hippocampal volume measurement appeared among top features for the model using imaging and proteomic features.

The importance of proteomic features shows a similar pattern. Apolipoproteins showed up frequently among the top features in both the model using proteomic features alone and the model using imaging and proteomic features. apolipoprotein E (APOE) and apolipoprotein ε4 (APOE ε4) have been previously studied as predictors for AD [Wolk].

ACKNOWLEDGMENTS

This research work was partially supported by the National Science Foundation EPSCOR DART grant, and the National Science Foundation with grant number 1452211, 1553680, and 1723529, National Institute of Health grant R01LM012601, as well as was partially supported by National Institute of Health grant from the National Institute of General Medical Sciences (P20GM103429).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Causey, Jason, et al. "Spatial Pyramid Pooling with 3d Convolution Improves Lung Cancer Detection." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, pp. 1–1., doi:10.1109/tcbb.2020.3027744. [@Causey]
2. Chen, Tianqi, and Carlos Guestrin. "XGBoost." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16, 2016, doi:10.1145/2939672.2939785. [@Chen]
3. "Datasets - Nlst - the Cancer Data Access System." *National Cancer Institute - Cancer Data Access System*, cdas.cancer.gov/datasets/nlst/. [@Datasets]
4. National Lung Screening Trial Research Team, Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, Galen B, Gareen IF, Gatsonis C, Goldin J, Gohagan JK, Hillman B, Jaffe C, Kramer BS, Lynch D, Marcus PM, Schnall M, Sullivan DC, Sullivan D, Zylak CJ. The National Lung Screening Trial: overview and study design. *Radiology*. 2011 Jan;258(1):243-53. doi: 10.1148/radiol.10091808. Epub 2010 Nov 2. PMID: 21045183; PMCID: PMC3009383. [@National]
5. Stubblefield, Jonathan, et al. "Cardiac or Infectious? Transfer Learning with Chest x-Rays for Er Patient Classification." 2020, doi:10.1101/2020.04.11.20062091. [@Stubblefield]
6. "Trial Summary - Learn - Nlst - the Cancer Data Access System." *National Cancer Institute - Cancer Data Access System*, cdas.cancer.gov/learn/nlst/trial-summary/. [@Trial]
7. Ahmed, Azeemuddin, and Mark A. Graber. "Evaluation of the Adult with Dyspnea in the Emergency Department." UpToDate, <https://www.uptodate.com/contents/evaluation-of-the-adult-with-dyspnea-in-the-emergency-department>. [@Ahmed]
8. Chen, Tianqi, and Carlos Guestrin. "XGBoost." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16, 2016, doi:10.1145/2939672.2939785. [@Chen]
9. Holubar, Marisa, and Stan Deresinski. "Antimicrobial Stewardship in Hospital Settings." UpToDate, <https://www.uptodate-com.ezproxy.library.astate.edu/contents/antimicrobialstewardship-in-hospital-settings>. [@Holubar]
10. Rajpurkar, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." ArXiv.org, 25 Dec. 2017, arxiv.org/abs/1711.05225. [@Rajpurkar]
11. Wang, Xiaosong, et al. "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi:10.1109/cvpr.2017.369. [@Wang]
12. Stubblefield, Jonathan, et al. "Cardiac or Infectious? Transfer Learning with Chest x-Rays for Er Patient Classification." 2020, doi:10.1101/2020.04.11.20062091. [@Stubblefield]
13. ADNI : about. ADNI. (n.d.). Retrieved October 20, 2021, from <http://adni.loni.usc.edu/about/>. [@about_adni]
14. ADNI : methods and Tools. ADNI. (n.d.). Retrieved October 20, 2021, from <http://adni.loni.usc.edu/methods/>. [@methods_adni]
15. Biomarkers consortium plasma proteomics data primer ... (n.d.). Retrieved October 20, 2021, from http://adni.loni.usc.edu/wp-content/uploads/2010/11/BC_Plasma_Proteomics_Data_Primer.pdf. [@biomarkers]
16. Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2939672.2939785> [@Chen]

17. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020). [[@Lundberg_Tree](#)]
18. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Advances in neural information processing systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017). [[@Lundberg](#)]
19. Lundberg, S. SHAP: SHapley Additive exPlanations. <https://github.com/slundberg/shap>. (2020). [[@Lundberg_Github](#)]
20. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, (2/1/2011), 2825–2830. [[@Pedregosa](#)]
21. Petersen, R. C., Yaffe, K., & Wilterdink, J. L. (2021). Mild cognitive impairment: Epidemiology, pathology, and clinical assessment. In *UpToDate*. essay, s.n. [[@Petersen](#)]
22. Wolk, D. A., Dickerson, B. C., DeKosky, S. T., & Wilterdink, J. L. (2021). Clinical features and diagnosis of Alzheimer disease. In *UpToDate*. essay, s.n. [[@Wolk](#)]